TP2 - Competencia Kaggle

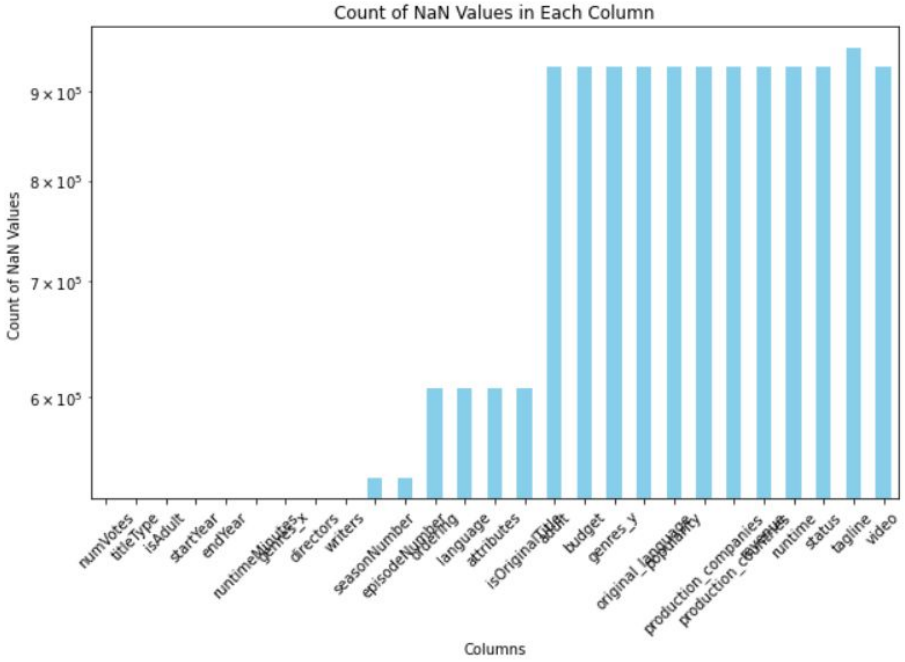
Alexander van Tol

Contenido

- Data set
- Modelo Baseline
- Selección modelos
- Descripción final de modelo
- Limitaciones y posibles mejoras

Data set

- Tamaño → (977541, 28)
- Predecir ‘averageRating’
- Limpieza de Nans



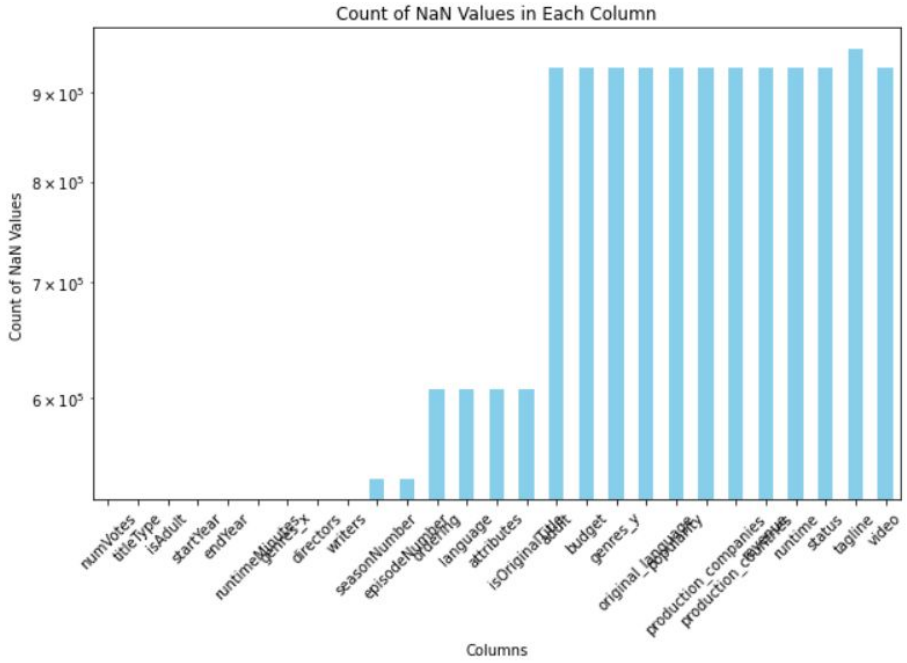
Modelo Baseline

avg_rating = df['averageRating']

df = df.drop('averageRating', axis = 1)

- Solo numérico

- Primeras siete columnas



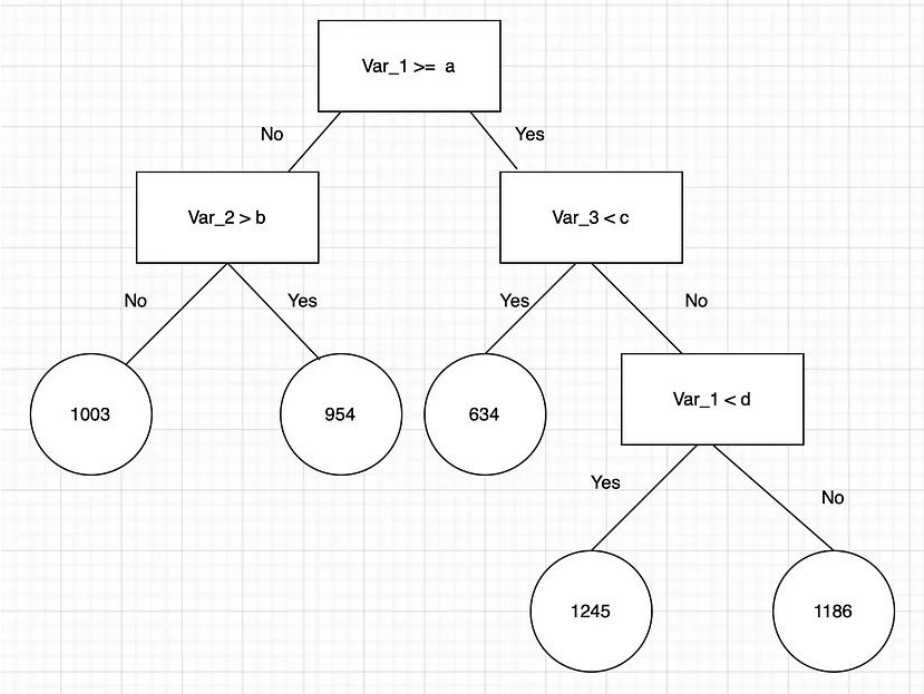
RF = RandomForestRegressor(n_estimators=50) → Accuracy: 20%

Modelo Baseline

Random Forest Regression:

- Combinación Decision Tree y Ensemble learning

- Bootstrapping algorithm



Selección Modelos

Probar otros modelos:

Regressión lineal → 4%

KNN → 3%

DecisionTreeRegressor → 15%

GradientBoostingRegressor → 13%

—----------

LabelEncoder() de titleType → 23%

Más preprocesamiento/aumentar tamaño de datos?

Modelo final

Nueva columna ‘Years’ → diferencia entre ‘startYear’ y ‘endYear’

Procesamiento de Nans en ‘episodeNumber’ y ‘seasonNumber’

	directors	writers
0	nm0883334	nm0844784,nm0305863
1	nm2291816,nm3088555,nm4930005,nm1746040	nm1707665,nm0789712,nm0403945,nm1826186,nm0630...
2	nm0414025	nm0414025,nm3692091,nm1620376
3	nm2977268	nm2977268,nm0415515
4	nm2366663	nm4290500,nm4289029

	directors	writers
0	0883334	0844784
1	2291816	1707665
2	0414025	0414025
3	2977268	2977268
4	2366663	4290500

Dummies columna para ‘titleType’ y ‘genres_x’ en vez de LabelEncoder()

Modelo final

num_6_col = ['numVotes','isAdult', 'Years', 'runtimeMinutes','seasonNumber','episodeNumber']

columns_to_exclude = ['language', 'adult', 'genres_y', 'original_language', 'production_companies', 'production_countries','attributes', 'status','tagline','video']

final.shape → (977539, 47)

RandomForestRegressor(n_estimators = 50) → ~49%

Modelo final

Selección de hiper parámetros manual:

→ test_train_split(stratify=avg_rating) para poder validar (val =20% de datos)

Tiempo de entrenamiento largo >10min

100 - 300 n_estimators → solo aumentó 1-2%

min_samples_split/min_samples_leaf → aumentó poco

Limitaciones y posibles mejoras

- Optimización de hiper parámetros

- Más variables a costa del tamaño de los datos? → relaciones sutiles

- Más pre-procesamiento de datos

- Reducir overfitting → max_depth = ~ , min_samples_split/leaf = ~